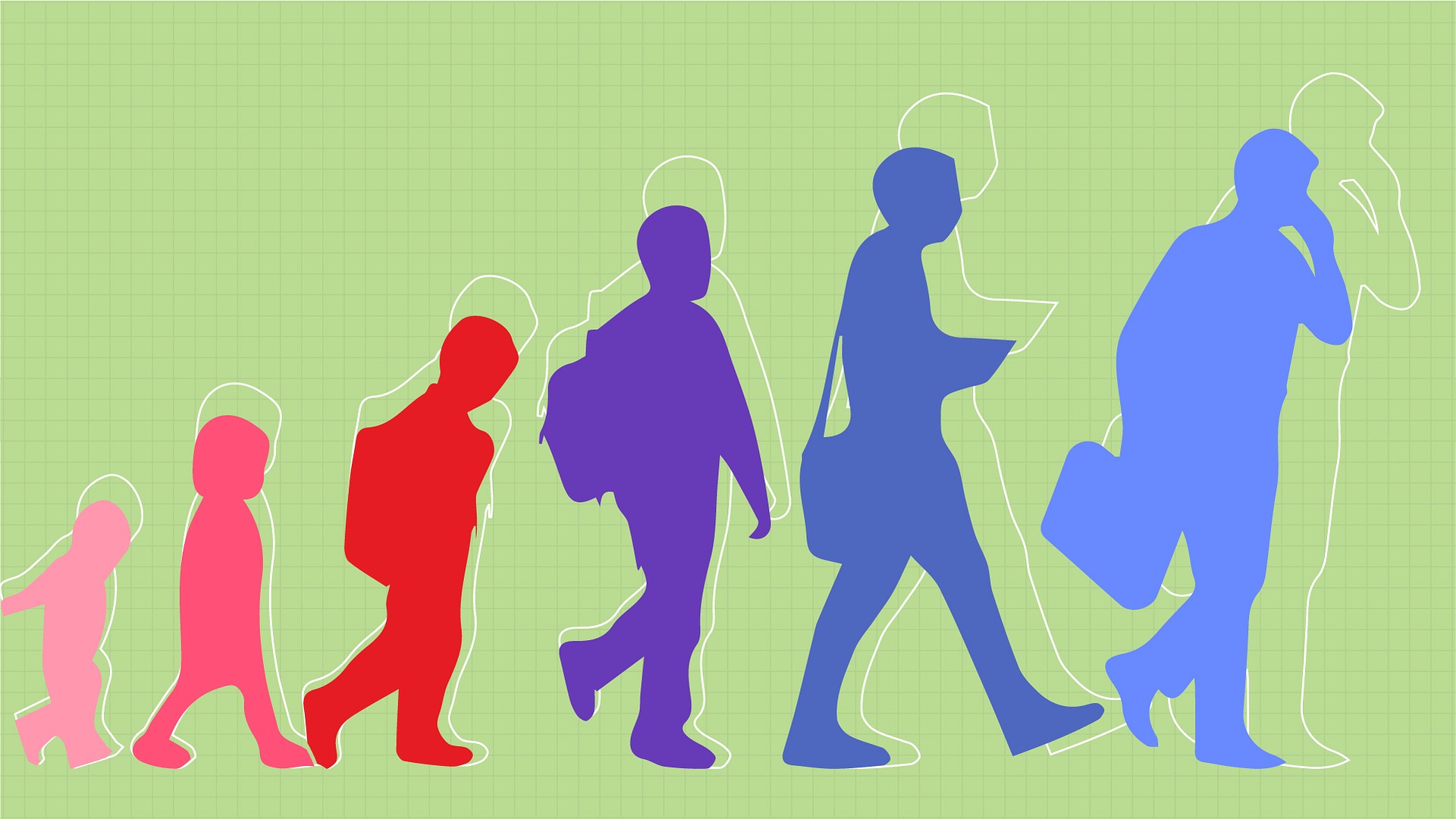
STRATEGIC THINKIG

CAPSTONE PROJECT

LIFE EXPECTANCYMachine learning prediction models



ADEWALE TAOFIQ ARIWO-OLA

SBS

**ABSTRACT**

This study explores the application of machine learning models to predict life expectancy based on various set of features such as Infant deaths, adults’ mortality, alcohol consumption, incidents of HIV, GDP per capital, schooling that are presented in a data set that was downloaded from dataset repository Kaggle. The dataset utilized for this research comprises a comprehensive collection of health-related attributes obtained from diverse populations of different countries. Exploratory Data Analysis and feature engineering techniques are applied to enhance model performance, and extensive preprocessing is conducted to handle imbalance class in target variable and outliers. The models are trained and validated using a robust set of evaluation metrics, including mean squared error, accuracy, Precision, F1-Score and Recall. Results demonstrate the efficiency of machine learning models in accurately predicting life expectancy, showcasing the potential for personalized health interventions and resource allocation. Additionally, interpretability analyses are conducted to elucidate the key features influencing the predictions, providing valuable insights for healthcare professionals and policymakers. The study also discusses challenges associated with implementing life expectancy prediction models in real-world healthcare settings. In conclusion, this research contributes to the growing body of knowledge in the field of predictive healthcare analytics, emphasizing the promising role of machine learning in advancing our understanding of life expectancy determinants.

# INTRODUCTION

The aim of the capstone project is to apply skills obtained during my learning for Higher Diploma Data Analytics for Business to evaluate a possibility to predict a chosen target variable based on other independent variables. I will use data exploratory analysis (EDA) and statistical methods to understand patterns and trends of a dataset and apply a predictor machine learning model on the processed data.

The data "Life Expectancy data" used for the initial phase of the capstone was downloaded from <https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who> and has been adjusted into a new dataset "Life Expectancy-Data-Updated"-<https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated>. Data contains life expectancy, health, immunization, and economic and demographic information about 179 countries from 2000-2015 years with 21 variables and 2864 rows. The database has one variable that categorizes countries into two groups: Developed vs Developing countries. Life expectancy is a statistical measure that represents the average number of years a person is expected to live based on various factors such as their birth year, gender, and other demographic characteristics. It is typically expressed as an average number of years and is often used as an indicator of the overall health and quality of life in a particular country or region.

The capstone project is executed using the CRISP DM methodology. As expected, changes happened through some stages of the project as I need to re-assess and re-evaluate some of these stages and come up with alternative means of achieving the project goals.

Life expectancy is a key summary measure of the health and wellbeing of a population. A nation’s life expectancy reflects its social and economic conditions and the quality of its public health infrastructure, among other factors. Monumental improvements in life expectancy have been the predominant trend for higher income, developed countries over the course of the 20th and 21st centuries (J*essica Y Ho and Arun S Hendi, 2018)*

The United Nations committee for Development Policy review the list of Low Developed Countries every three years and make recommendations on the inclusion and graduation of eligible countries using the following criteria.

1. Income: Measured by the Gross National Income (GNI) per capita

GNI per capita provides information on the income status and the overall level of resources available to a country.

1. Human Assets: Measured by the Human Asset Index (HAI)

The HAI is a measure of level of human capital. Low levels of human assets indicate major structural impediments to sustainable development and a lower HAI represents a lower development of human capital (Fig. 1)

1. Economic and Environmental Vulnerability: Measured by the Economic and Environmental Vulnerability Index (EVI)

The EVI is a measure of structural vulnerability to economic and environmental shocks. High vulnerability indicates major structural impediments to sustainable development. A higher EVI represents a higher economic vulnerability (Fig.2)

Fig.1: Human Asset Index (HAI)

A picture containing diagram

Description automatically generated

Fig.2: Economic and Environmental Vulnerability Index (EVI)

Diagram

Description automatically generated

Source: UN DESA (https://www.un.org/development/desa/dpad/least-developed-country-category/ldc-criteria.html?target=human-assets)

# HYPOTHESIS

Can socioeconomic, health, environmental and geographical location criteria be used to predict if a populations life expectancy is low or high?

# GENERAL GOAL

The goal of this project is to analyze the dataset and generate a synthetic data to be used to construct a machine learning model that can predict life expectancy in regions of countries around the world based on factors that are represented the dataset.

# DEFINING THE PROBLEM STATEMENT

I am creating a categorical machine model which can predict life expectancy of a population in regions of countries around the world.

Predictors: Country, Region, Year, Infant deaths, Under-five deaths, Adult mortality, Alcohol consumption, Hepatitis B, Measles, BMI, Polio, Diphtheria, Incidents HIV, GDP\_per\_capita, Population\_mln, Thinness\_ten\_nineteen\_years, Thinness\_five\_nine\_years, Schooling, Economy\_status\_Developed, Economy\_status\_Developing,

Target Variable: Life Expectancy Categpry

Low Life Expectancy Category = 0

High Life Expectancy Category = 1

# SUCCESS CRITERIA/INDICATORS

I was able to establish correlation between life expectancy and chosen socioeconomic, geographical location and health conditions.

Random Forest Model was used to predict if a populations life expectancy is low or high with an outcome of 100% accuracy.

Independent features used for predicting country status for the machine learning model were extracted based on their importance.

# DATA SOURCE

The data was obtained from a data set repository Kaggle "Life Expectancy-Data-Updated"-<https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated>

# TECHNOLOGIES USED

There are certain set of Python libraries that is collectively labeled the Scientific Stacks. This stack comprises, among others, the following:

NumPy: NumPy provides a multidimensional array object to store homogenous or heterogenous data; it also provides optimized functions/methods to operate on this array object.

Matplotlib: This is the most popular plotting and visualization library for python, providing both 2D and 3D visualization capabilities.

Pandas: Pandas builds on NumPy and provides richer classes for the management and analysis of time series and tabular data; it is tightly integrated with matplotlib for plotting and PyTable for data storage and retrieval. (Python for Finance. Analyze Big Financial Data, Yves Saarland, 2014.Pg 16 & 17)

Seaborn is a python library built on top of matplotlib and allows you to easily produce prettier (and more complex) visualization (Data Science from Scratch. Joel Grus,2015. Pg.83)

Also, we are importing machine learning library for Python such as Scikit-learn, often called sklearn. T provides simple and efficient tools for data analysis and modelling, including various machine learning algorithms for classification, regression, clustering, dimensionality reduction, and more.

SciPy is an open-source library for mathematics, science, and engineering. It is built on top pf NumPy library and provides additional functionality for wide range of scientific and engineering applications.

Machine Learning model were built using Decision Tree and Random Forest Classifier algorithms which builds multiple decision trees and combines their predictions to improve the accuracy and reduce overfitting.

# EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis (EDA) is a useful method generally employed to understand data set by summarizing main characteristics of the data set and often visualizing by plotting.

Making informative visualizations (sometimes called plots) is one of the most important tasks in data analysis. It may be a part of the exploratory process-for example, to help identify outliers or needed for data transformations, or as a way of generating ideas for models. (Wes McKinney, 2017)

Python has become a component of the much broader Jupyter open-source project, which provides a productive environment for interactive and exploratory computing. Its oldest and simplest mode is as an enhanced python shell designed to accelerate the writing, testing and debugging of Python code. IPython system can be used through Jupyter Notebook, an interactive web-based code “notebook” offering support for dozens of programming languages. The IPython shell and Jupyter notebooks are especially useful for data exploration and visualization. (Wes McKinney, 2017)

All data analysis and visualization for completing this project are generated using python.

Python contains lots of data science libraries, frameworks, modules, and toolkits that efficiently implement the most common as well as least common data science algorithm and techniques. (Joel Grus, 2015)

# DATA UNDERSTANDING

The dataset has 2864 entries in 21 columns of which float64(11), int64(8), object (2). There are neither missing values nor duplicated rows in the dataset.

# DATA PREPARATION And PROCESSING

In this phase, the dataset will be converted into meaningful and useful information. This involves series of operation that manipulates, analyze, and transform data to extract insight, make decisions or generate reports. For the dataset some of the aspect of data processing phase used are data cleaning, data transformation, data analysis and data visualization.

To prepare our dataset for insightful analysis and importing it to a machine learning models following steps were taken:

* Number of variables (columns) was reduced from 21 to 19 by dropping features ‘Year\* and ‘Countries’ not significant to the project goal.
* Features in the target variable (Life Expectancy) columns was categorized into ‘Low = below 64’ or ‘High = over 64’. The 2020 average normal retirement age across OECD countries for an individual with a full career and who entered the labor market at age 22 was equal to 63.4 years for women and 64.2 years for men.
* A new column was created for the target variable ‘Life expectancy category’ and text classification of the target variable was replaced with 0 and 1 for each category of low and high respectively.
* The dataset was analyzed for duplicates, but no duplicated rows identified.
* Data in all columns, apart from "Life expectancy" was converted to floats to ensure it can be used with a wide range of models without any compatibility issue.
* Data was shuffled and scaled using Standardization algorithm StandardScaler. Data standardization is an important technique in data preprocessing. The goal of data standardization is to transform the numeric variables so that each variable has zero mean and unit variance.

## **BASIC DATA EXPLORATORATION**

In this phase, I am gauging the overall data. The volume of data and the types of columns present in the data.

This initial assessment of the data is done to identify which columns are Quantitative, Categorical or Qualitative.

All data analysis and visualization are generated using python.

Python contains lots of data science libraries, frameworks, modules, and toolkits that efficiently implement the most common as well as least common data science algorithm and techniques. (Joel Grus, 2015)

This step helps to start the column rejection process. I am looking at each column carefully and ask, does this column affect the values of the Target variable? In this case I am asking the question does this column affect the life expectancy of a population? If the answer is a clear "No", then I will remove the column immediately from the data, otherwise I am keeping the column for further analysis.

There are four commands which are used for Basic data exploration in Python.

head() : This helps to see a few sample rows of the data

info() : This provides the summarized information of the data

describe() : This provides the descriptive statistical details of the data

nunique(): This helps us to identify if a column is categorical or continuous

Basic data exploratory analysis shows the data contains 21 columns of variables and 2864 rows of observations with categorical and numerical datapoints. There are no missing values in all columns and rows are not duplicated in the dataset.

Fig. 1: Features in columns and data types.

Using the nunique() command to find unique values for each column to understand which column is categorical and which one is Continuous, Typically if the number of unique values are less than 20 then the variable is likely to be a category and numbers greater than 20 are likely to represent continuous variables.

Fig.2: Unique value for each column

### **LOOKING AT THE DESCRIPTIVE STATISTICS OF THE DATASET**

Descriptive measures that indicate where the center or most typical value of a data set lies are called measure of central tendency or more simply, measure of center. Measure of centers are often called averages. There are 3 most important measure of center: the mean, median and mode. (Neil A. Weiss (Introductory Statistics, 2017, pg.118)

In Fig 3, we can see the mean of the data showing average of each of the numerical values, Also, standard deviation of each column showing how values varies. This is a measure of dispersion from the average value.

Also displayed are the lower 25% values, average 50% values and upper 75% values representing first quartile (Q1), second quartile (Q2) and thirst quartile (Q3) respectively.

The standard deviation measure variation by indicating how far, on average, the observation is from the mean. For data set with large amount of variation, the observation will, on average, be far from the mean; so, the standard deviation will be large. For a data set with a small amount of variation, the observation will, on average, be close to the mean; so, the standard deviation will be small (Neil A. Weiss, Introductory Statistics, 2017 pg. 129)

There are inconsistencies in the min and max values in the data. The max and min values of one feature are significantly larger than the other features. These features are standardized using StandardScaler.

The RobustScaler ensures statistical properties for each feature that guarantee that they are on the same scale by using the median and quartiles. This makes the RobustScaler ignore data points that are very different from the rest (like measurement error). ( Andreas C. Müller & Sarah Guido,2017)